Comparing the Power of Plot Designs to Reveal Correlation

A thesis submitted in partial fulfilment of the requirements for the degree of:

Bachelor of Commerce (Honours)

In

Econometrics

Monash University
Faculty of Business and Economics

Written by Nathaniel Tomasetti
Supervised by Dianne Cook

October 2015

Comparing the Power of Plot Designs to Reveal Correlation

Nathaniel Tomasetti and Di Cook

Abstract – Visual inference in EDA is prone to type 1 errors from the over-interpretation of randomness [4, 6]. Two competing plot designs, the scatter plot and overlaid line graph are both popular in the analysis of time series data. Lineups [2, 10, 11, 16, 27] allow the visual inference power of a graphic display to be evaluated, and were used to compare the plot designs. We collected data on the detection rate of correlated pairs of AR(1) simulations, the time required and the confidence of the decision for 2088 lineup evaluations. The results show that the scatter plot is both the faster and more powerful plot design, despite its inability to display the time dimension.

Index Terms – Lineups, visual inference, power comparison, scatter plot, line graph, data visualisation, visual analytics, time series, temporal data

1. Introduction

In order to work with data, it first must be understood. Statistical inference requires hypotheses to be established prior to data collection, but often data is collected first. This is especially so today, for vast databases that have been assembled in the big data era, that now need the data scientist to unravel the meaning of the numbers. Without preset hypotheses to test, the power of statistical inference is impotent, and without hypotheses the data analyst can stumble blindly trying to build up models of structure in the data. To understand data requires good visualisation. This idea was formed as early as the 18th century, when William Playfair institutionalized the then revolutionary idea of graphing government and economic data. Far easier than reading tables of numbers, these ideas were powerful, and by providing the basic building blocks for plotting statistical data, his graphic designs became a conduit for the communication of otherwise complex information [20]. Since then, advances in computing power have allowed statistical graphics to flourish, spearheaded by Tukey in 1965 into the new domain of exploratory data analysis (EDA) [7, 23]. EDA can be thought of as a well understood [3, 25] set of tools and techniques required to visualise information, to physically see what the data contains. In particular, it incorporates a free roaming approach, where the analyst is able to explore structure to find whatever relationships and structure exists within. Critically, the analyst does not have to have any pre-conceived ideas or hypotheses -- they are not specifically looking for any one particular thing. EDA emphasises letting the data inform us and can lead to the discovery of otherwise unexpected relationships, many of which may seem to become completely obvious after discovery. With the knowledge provided by EDA, ideas are generated about what relationships between variables may potentially exist. This then enables the analyst to use these new hypotheses upon which to apply classical inference and rigorously conduct tests with new data. EDA is also related to the field of model diagnostics (MD), where a model can be continuously refined through the visualisation of its fit, its residuals, and the interactions with its variables. Both EDA and MD follow the same framework: Visualise the data, look for patterns that suggest an underlying relationship, and if one is found implement it into the model then continue the exploration of the data. It has long been thought that EDA and statistical inference were worlds apart, but recent work [2, 16] bridges this chasm. Framing a data plot as a test statistic, which when compared to plots of null data, places EDA into the statistical inference recipe.

The null hypothesis underlying a particular plot, is generically that there is no pattern, and particular types of plots implicitly regulate what "no pattern" means. For example, a scatter plot of two variables is used to explore for some sort of association, so the implicit hypothesis is that there is no association between the two variables. The alternative hypothesis is that there is some association, although it is not required to specify precisely the type of association. The plot of the data is placed in a lineup of plots of null data, data generated assuming that the null hypothesis is true. If an impartial observer asked to pick the plot that is different from the rest, picks the data plot this suggests there exists a pattern that is not the result of chance, and the null hypothesis is rejected. For a scatter plot, this departure from the null, might be a single outlier, or few outliers, a non-linear pattern, or clusters, which the human eye detects as more different in this data plot than in any of the nulls. This is the reason why visualisation remains important today, human eyes can detect patterns which would not be detected mathematically. But eyes need calibration, which the lineup protocol provides.

On its own, with a single plot, because of random sampling in collecting data, it is easy for the analyst to imagine a pattern when no real structure exists in the population. Visual skills of an observer in EDA is prone to Type I error, where the null hypothesis is rejected when it is actually true, caused by the inherently random formation of patterns when visualised which are attributed to structure rather than chance. Daniel [4] warns against this, by providing 40 pages of plots from the null distribution, he encourages data analysts to understand the patterns that can be created from data without inherent structure. (This is what Buja et al [2] call the Rorschach protocol.) By being aware of what can appear in this type of data, the analyst should be more wary of claiming any visual feature they find is a true relationship. But this is not enough, even seasoned data analysts can be misled. Diaconis [6] introduced the notion of 'magical thinking' which argues that people commonly suffer from the over-interpretation of randomness, particularly if it matches a pre-conception. If it suits their particular bias, an analyst may make false discoveries of structure and false rejections of the null hypothesis. Whilst using both can complement each other, the treatment of Type I error has led EDA and inferential statistics to be considered as very disparate pursuits. The mathematical rigor that governs classical inference relies on a framework that acknowledges and controls for the rate of Type I error. Section 2 describes the visual inference protocols which put EDA more firmly into the rigorous framework of statistical inference.

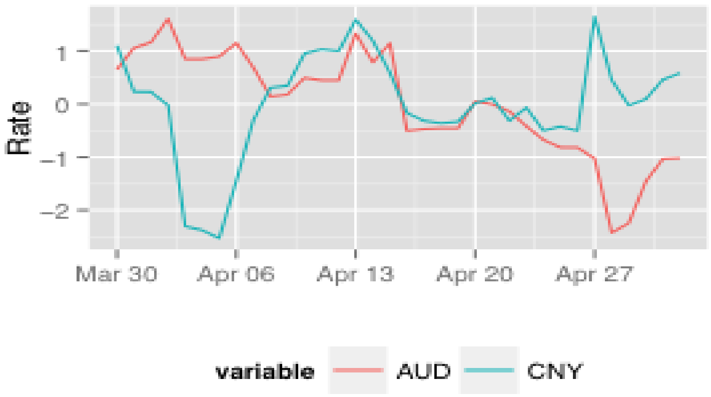


Figure 1: Exchange rates for Australian dollars and Chinese yuan against the United States dollar from Mar 30-May 2, 2015, shown as an overlaid line graph. (Both sets of values were standardized to bring them to the same scale for plotting.) Do you think a relationship exists between the two?

In this research we utilise the methods in visual inference to study plot design to investigate the relationship between two temporal variables. There is a substantial literature suggesting that scatter plots are the appropriate display to read association [3, 8, 14], however, when the two variables are temporal it is common to display them as time series, drawn on the same plot. Temporal dependence is present in many macroeconomic variables [16, 17] and there is a strong argument for its presence in financial data [21, 24]. It is a common, but untested, belief that the inclusion of time in the graphics will increase the detail of information displayed and allow relationships to be examined more comprehensively, so many analysts utilise the overlaid line graph to examine their data.

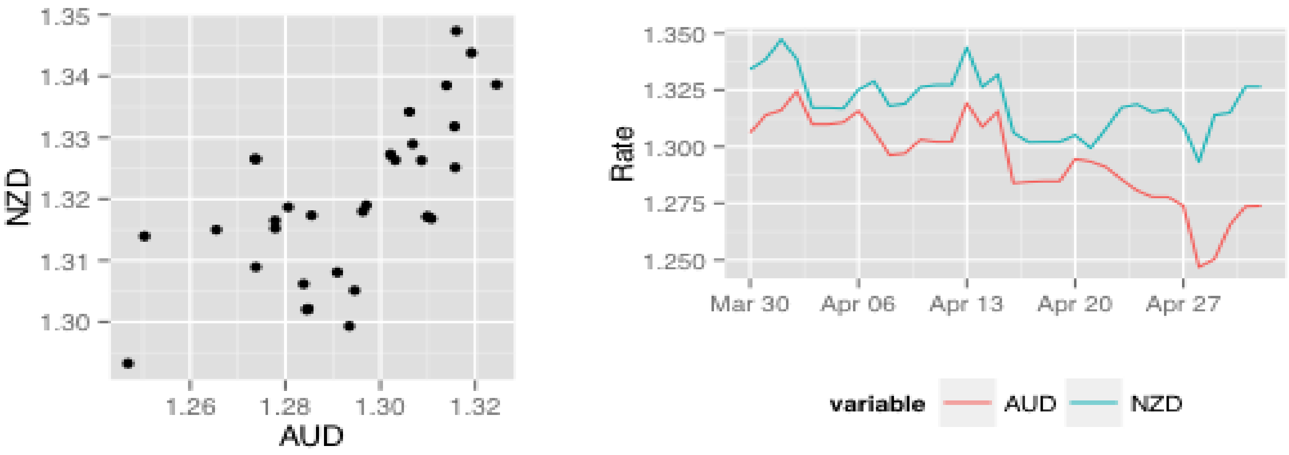


Figure 2: Exchange rates for Australian and NZ dollars against US dollar, from Mar 30-May 2, 2015, shown using a scatter plot (left) and overlaid line graph (right). Which one can you read the correlation from more accurately? Point-wise linear correlation between the two series, ignoring autodependence, is 0.64.

To decide on an appropriate display requires awareness of cognitive principles in psychology such as the Gestalt law of common fate described by Wertheimer in 1923 [22, 26]. This law describes a tendency in human perception to 'see' a positive relationship between two objects that briefly move together over time, leading people to believe that they stay linked and share a 'common fate' (Figure 1), even if no relationship exists. Lee and Blake's [13] work finds that humans are able to perceive a relationship between objects that move with temporal synchrony in accordance with the Gestalt Law, implying that a graphic than can utilise a dimension for time may be able to use the information provided by temporal dependence and potentially outperform the graph types that do not, such as the scatter plot [18]. However, keep in mind magical thinking [6], the desire to find any pattern may lead to a false discovery of positive correlation. The law then has two effects, if positive correlation exists in temporal data, the observer may be drawn to it easier. If positive correlation does not exist, the observer may be fooled into thinking it does anyway. Either effect would lead to a more confident rejection of the null than if the same data was displayed as a scatter plot. Figure 1 highlights the dangerous interaction between 'magical thinking' and the Gestalt Law, the observer is drawn to the period from April 8 to April 22, and may falsely believe that the two currencies may continue to move together. However, over the entire series the linear correlation between the two currencies is -0.12, indicating there is no actual relationship. Incorrectly relying on two currencies, or many other sets of economic variables, being dependent on each other can be extremely dangerous for a firm. Work by Robbins [19] is similarly critical of the time-based line graph, where it is argued that information from overlaid line graphs can easily be misleading (Figure 3). When both lines dive after 1993, it is natural to compare the horizontal distance and say that the difference between the lines in 1994 is small. However, this is incorrect, as looking at the small horizontal distance is comparing the lines at different dates. It is not obvious that the difference in 1994 is the second greatest difference across the entire range, only slightly smaller than in 1993. Vanderplas [25] examines this effect in more detail. It is dubbed the 'Sine Illusion', where the human eye is poor at interpreting lines with such a sharp slope. As these problems are unique to the line graph, they argue that the scatter plot, which does not have any ability to display time, is the superior choice. However, in practice many data analysts are split between the two major alternatives for temporal data. The overlaid line graph is justified by its ability to present more information to the analyst; but many argue that it is this extra information that is misleading, and revert to the scatter plot for its strong non-temporal performance.

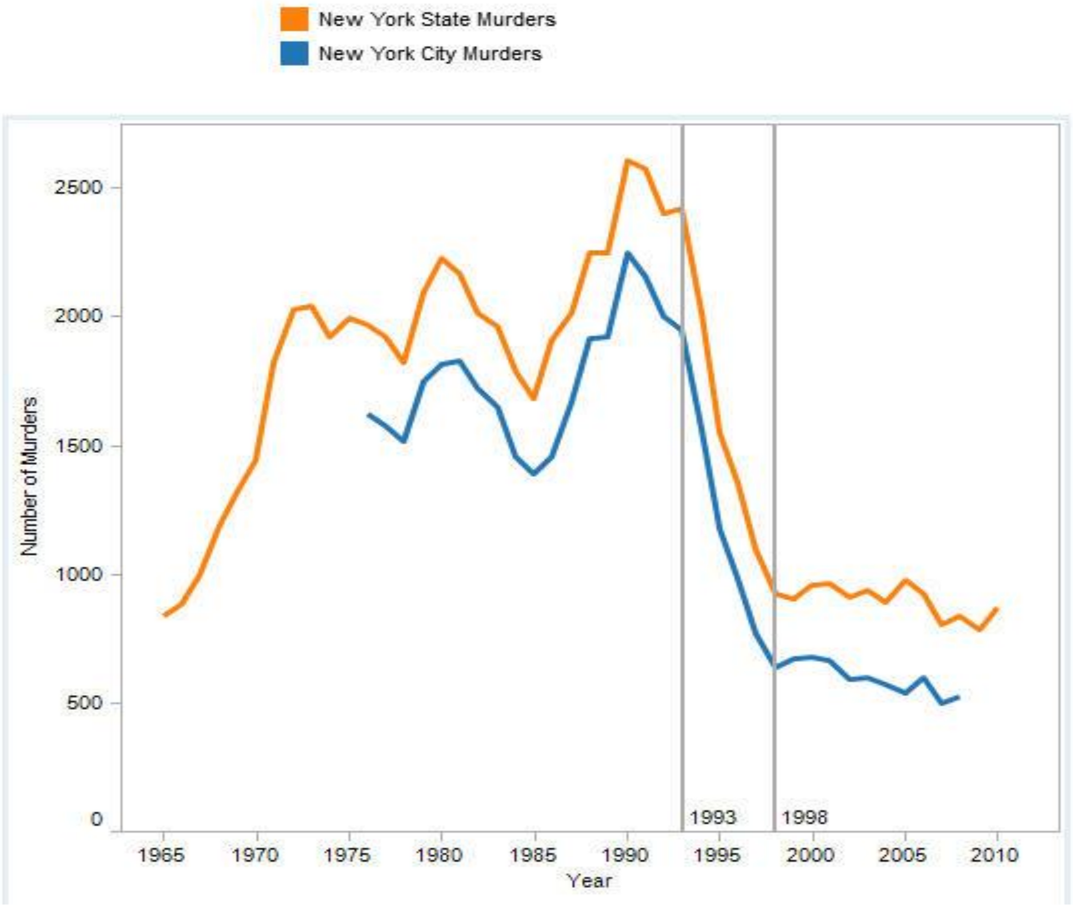


Figure 3 [19]: Murders in New York State (Orange) and New York City (Blue). The difference between the two represents murders outside of the city. What size would you judge the difference to be in 1994?

There has been other research on the perception of temporal displays, e.g. Javed et al. [12], but they do not examine perception of association. Javed et al. examined optimal ways to visualise local features, such as, which variable had the highest value at a given point in time, and global features, such as, comparing the size of the overall slope of many different variables. The finding was that a different graphical layout is optimal for each type of task. However, they did not test for the perception of correlation, which can be treated as both a local and a global feature within the graph. The slope of each variable must be compared at a point in time, and they must have a persistent relationship across at least the majority of the series. However there is a quandary, the visual features that improve local tasks are unsuited to global tasks and vice versa; so we must now find a form that is well suited to assess both types of tasks simultaneously. This paper addresses this deficiency. Section 2 describes the lineup method utilised to rigorously compare plot designs. Section 3 explains the experimental design. Section 4 contains the results of the experiment and Section 5 discusses the implications of the findings.

2. The Lineup Protocol

With the general lack of inference present in exploratory data analysis, how do we attempt to control error in hypothesis testing? In order to use visualisation effectively, we will get the most benefit if the type of graphic display chosen is best suited to the task at hand, be it EDA, MD, or even the presentation of results. We then need to find the statistical power of graphics, the ability of a graphical display to convey information on the structure of the data within. This can be found by using the lineup protocol developed in 2009 by Buja et al. [2], which is easily implemented with the nullabor R package. They created a statistically rigorous framework with properties explored primarily in Majumder et. al [16] and further in Hofmann et. al [11] that allows us to conduct these hypothesis tests with visual inference, thus to some degree allowing the conjoining of EDA with classical inferential statistics.

The lineup protocol is inspired in part by the police lineup [27]. We place the plot of the 'true' data generated with some underlying structure (The criminal) in amongst plots of data that were generated from a null distribution (The innocent people). If the real data plot is selected by an uninvolved observer as being most different from the other plots there is evidence that the structure of that data has led to a significant difference in the plot (That the criminal is sufficiently different from the innocents). This constitutes a rejection of the null hypothesis. If one of the null plots is chosen instead then either the plot design did not have the sufficient power to display the true relationship (A Type 2 error in EDA), or the null plot exhibited a strong relationship that was generated randomly and an analyst that saw this plot and decided that the data had some structure would've committed a Type 1 error. p_i, the probability that plot i is chosen in the lineup depends not only on the plot design d, but also the signal strength, q_i, of that plot and of every other competing plot in the lineup. It can be defined as some unknown function $f_{i,d}(q_1, \dots, q_{20})$ [11]. If the true plot is detected, it indicates that the plot design could convey the desired information about the underlying structure and that the true plot has a greater signal strength than those generated under the null distribution. The plot design that has a human observer selecting the true plot the most often will thus minimise both Type 1 error and Type 2 error in EDA hypothesis testing.

However, there is the possibility that the true plot was picked by chance, that we have committed a Type 1 error in the lineup test. For a lineup of m plots, the probability of selecting any plot when they are all generated by the null distribution is $1/m$, setting the Type 1 error rate, α, of a lineup hypothesis test to equal $1/m$. To control our error rate, it is initially obvious that we can simply change m, with an increase in null lineups having an inverse reduction in m. We recruit human observers to judge the lineups, but this can quickly lead to a large cognitive burden to sort through more and more null plots. A much more powerful option is to have multiple different viewers of each lineup, with K observers; the probability that at a particular plot was picked at least x times under the null hypothesis is binomially distributed [2, 11] with:

$$\mathrm{p-value} = \mathrm{P}(\mathrm{X} \geq \mathrm{x}|\mathrm{H}_0) = 1 - \mathrm{B}_{\mathrm{K},1/\mathrm{m}}(\mathrm{x}-1) \qquad (1)$$

If all K observers pick the same plot out of a lineup of twenty, we result in a p-value as small as $(1/m)^K$. This research used $m = 20$, giving us a significance level (and Type 1 error rate) of $\alpha = 0.05$. The plot of 'true' data is placed randomly amongst 19 null plots to form the lineup.

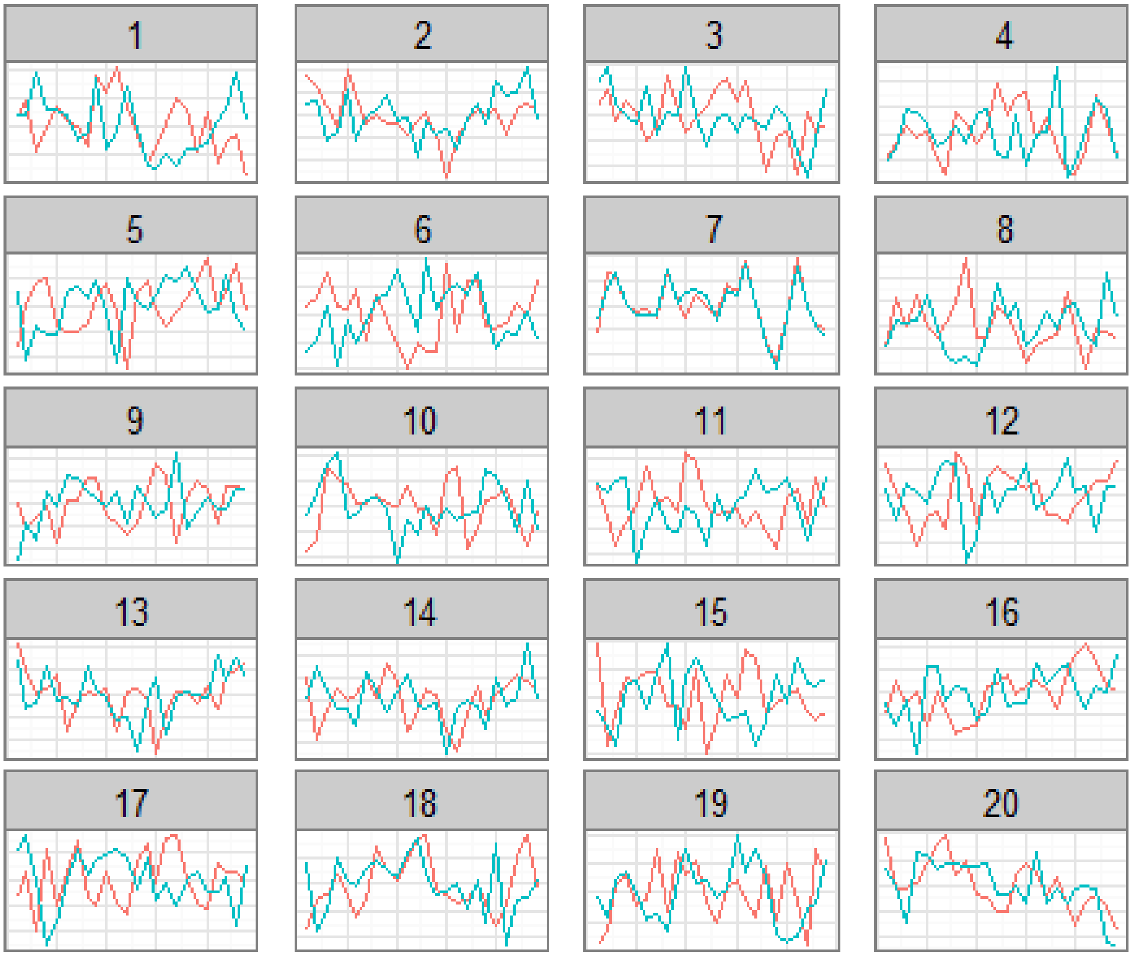


Figure 4: A lineup (m = 20) of line graphs.
Which plot shows the series with the strongest association?

Essentially, the more time our subjects pick the correct plot out of the lineup, the more confident we are that the real plot was visually significantly different to the nulls; and that the type of graphic involved has the power to display association. The lineup effectively allows us to conduct a hypothesis test on visual features, with the competing hypotheses:

H_0 : There is no association between the two series (evidence for H_0: the associated data is indistinguishable from the nulls).

H_1: There is an association between the two series (evidence against H_0: the associated data can be distinguished from the nulls).

As the power of a statistical test is defined as the probability of rejecting H_0 when it is false, the power of a lineup test is viewed as the probability of detecting the true plot. We approximate the power as $\hat{\pi} = x/K$, where x observers out of K correctly picked the true plot out of the lineup. We can them estimate the power difference of competing lineups, $\hat{\pi}_1 - \hat{\pi}_2$. An α × 100% confidence interval is calculated as:

$$\hat{\pi}_1 - \hat{\pi}_2 \pm \mathrm{t}_{1-\alpha,2\mathrm{n}-1}\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/\mathrm{n}_1 + \hat{\pi}_2(1-\hat{\pi}_2)/\mathrm{n}_2} \qquad (2)$$

Where $\hat{\pi}_\mathrm{i} = (\mathrm{x_i} + 1)/(\mathrm{n_i} + 1)$, x is the number of times a true lineup was correctly identified and n is the Welch-Satterwaite estimate for the degrees of freedom [10].

Some individuals may have a better natural ability at detecting the correct correlation pattern in the graphics, but by having each participant viewing multiple lineups and each lineup being viewed by multiple different participants, this can be controlled via a random effect variable in the model. We recruited participants through Amazon's Mechanical Turk, where the lineup protocol has been applied to a number of problems in prior papers [9, 16, 27]. Amazon’s Mechanical Turk [1] (MTurk) is a labour crowd-sourcing platform developed to give easy access to workers with basic tasks paid in line with the United States minimum wage. MTurk can be used to recruit subjects to read a variety of graphics and report on particular visual tasks. The time taken to decide, confidence in the decision and the reason for that decision can also be recorded. Heer and Bostock [9] use MTurk to replicate previous findings in graphical perception [3] and find that it is a valid method of data collection, once controls to eliminate anyone 'gaming' the system by randomly selecting answers to minimise time spent working are implemented. Further details of our use of Amazons’ Mechanical Turk is included in Section 3. The protocol has successfully been used to measure statistical power as a means to determine plot type superiority in Hofmann et al [10], and this work follows that approach.

3. Experimental Design

To allow for a greater control of distributions of the data in the lineups, we use simulated data. The data must contain autocorrelation for there to be meaningful information in the time dimension. To fulfil this requirement, each plot has a standardised pair of AR(1) models defined by:

$$Y_{1,t} = \beta \cdot Y_{1,t-1} + e_{1,t}$$
$$Y_{2,t} = \beta \cdot Y_{2,t-1} + e_{2,t}$$

With $t \in \{12,24,48,96\}$ and e_t being generated with from a bivariate normal distribution:

$$N\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

The variation in the covariance parameter ρ allows for control in the correlation in the pairs of data. The null data has no structure, so uncorrelated pairs are produced by setting $\rho = 0$. True data pairs are generated with $\rho = \pm\{0.3,0.5,0.7,0.9\}$. Pilot testing found that the rate of picking the true data plot fell to approximately 5%, the $1/m$ rate of randomly picking the true plot, for pairs with correlation below $\rho = 0.3$. Further to this, relationships weaker than this point are often of little interest to analysts. Hence we found it unnecessary to test a more full range of correlations. Simulating data does not produce correlations exactly equal to ρ, especially for small samples. Simulated data was accepted as real data if pairwise correlation between the two series is within 0.015 of the desired value of ρ. The null pairs often had correlation generated spuriously through the simulation process. This was particularly problematic for generating small sample data, but setting $\beta = t/100$, to a maximum of 0.5, for both true and null data reduced spurious correlation to manageable levels.

For particularly large time series, the AR(1) model can create particularly 'jagged' lines that may be unrealistic for many actual applications. To counter this, additional sets of data were created with a Hodrick-Prescott filter to capture the trend of the model and remove the noise in the cycle component. This filter creates a trend by penalising both errors in smoothness and in fit. The trend of the series, τ_t is defined by the equation:

$$\min_\tau\left(\sum_{t=1}^{T}(y_t - \tau_t)^2 + \lambda\sum_{t=2}^{T-1}[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})]^2\right)$$

Setting $\lambda = 1$ produced a smoothing that is more consistent with real world time series applications and did not introduce excessive spurious null correlation. However, for $t = 12$, 24, any form of smoothing was unfeasible without introducing extreme null correlation, hence smoothing was only used for $t = 48$, 96. Each generated lineup was produced as both a scatter plot and an overlaid line graph.

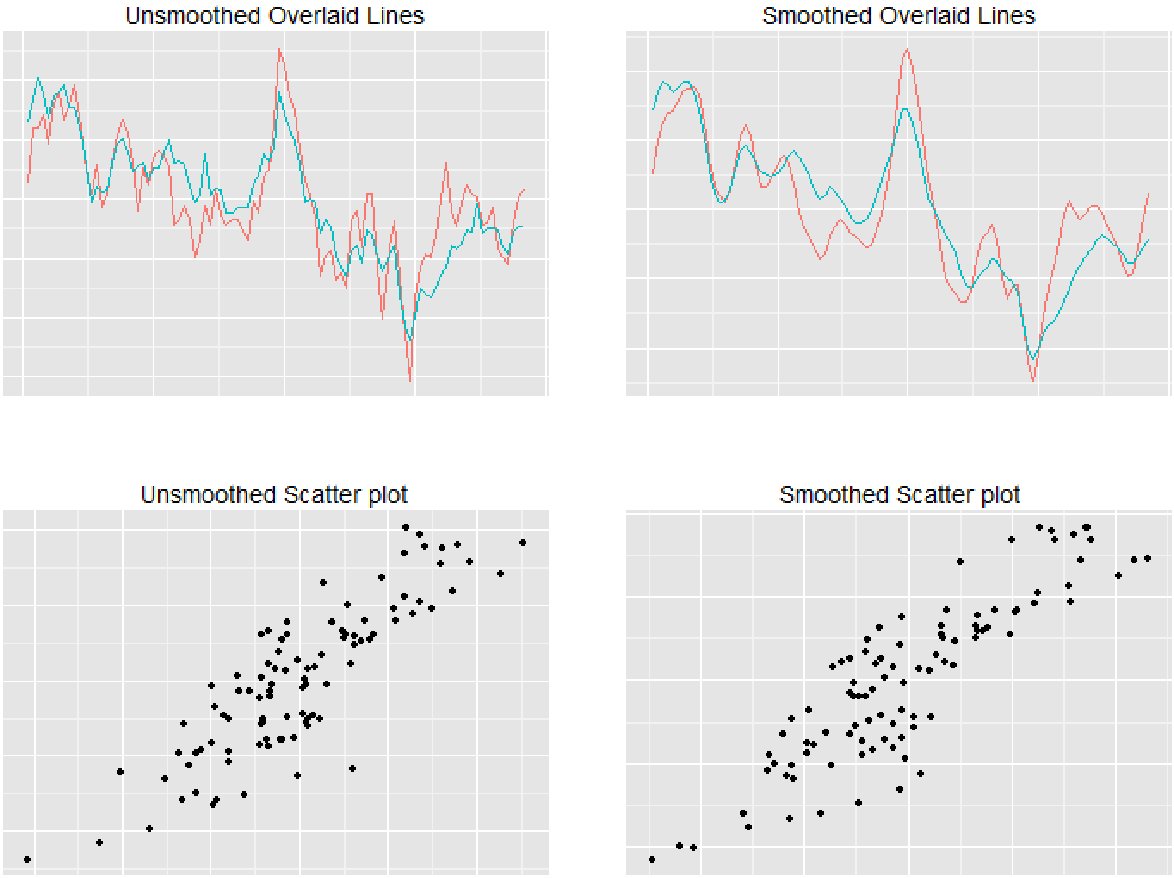


Figure 5: The same AR(1) data both smoothed and unsmoothed. t = 96, correlation = 0.86

In this research, we had three basic factors:

- Factor 1: Plot design, levels = scatter plot, overlaid time series
- Factor 2: Sample size, levels = 12, 24, 48, 96
- Factor 3: True plot correlation, levels = $\pm\{0.3, 0.5, 0.7, 0.9\}$

Additionally, there was a fourth factor for t = 48, 96:

- Factor 4: Smoothed, levels = yes, no

Each combination of factors was replicated in three different lineups, giving a total of 288 lineups (2 plots x 8 correlations x 6 sample sizes & smoothness combinations).

The position of the actual data plot in the lineup was randomized. The order that the factors are presented to a subject were also randomized, and each subject saw 5 scatter plots and 5 overlaid lines at a range of all absolute correlation levels. Each subject did not see the same data more than once. Subjects were asked to pick the plot in the lineup that has the strongest association, with basic examples of negatively and positively correlated plots. They also answered with reasons for choosing the plot, and the confidence that they have that this really is the plot showing the strongest correlation. Each subject evaluated 10 lineups, plus a further two trials to eliminate people attempted to game the system. If a trial was not answered correctly, the subject could not evaluate the remaining lineups. If it was answered correctly, the trial data was discarded and the subject moved onto the rest of the lineups. Earlier work [15] has found that repeated evaluations from a subject does not increase their ability to detect the true plot from a lineup, justifying our treatment of the results as independent evaluations. We recruited subjects from Amazon's Mechanical Turk service and ensured that each lineup was viewed multiple times. In total there were 2088 lineup evaluations, however to better control each subject’s individual visual ability we removed data from subjects with less than ten individual evaluations leaving us with 1684 evaluated lineups for the analysis. The study can be freely trialed at http://104.236.245.153:8080/mahbub/turk18/index.html

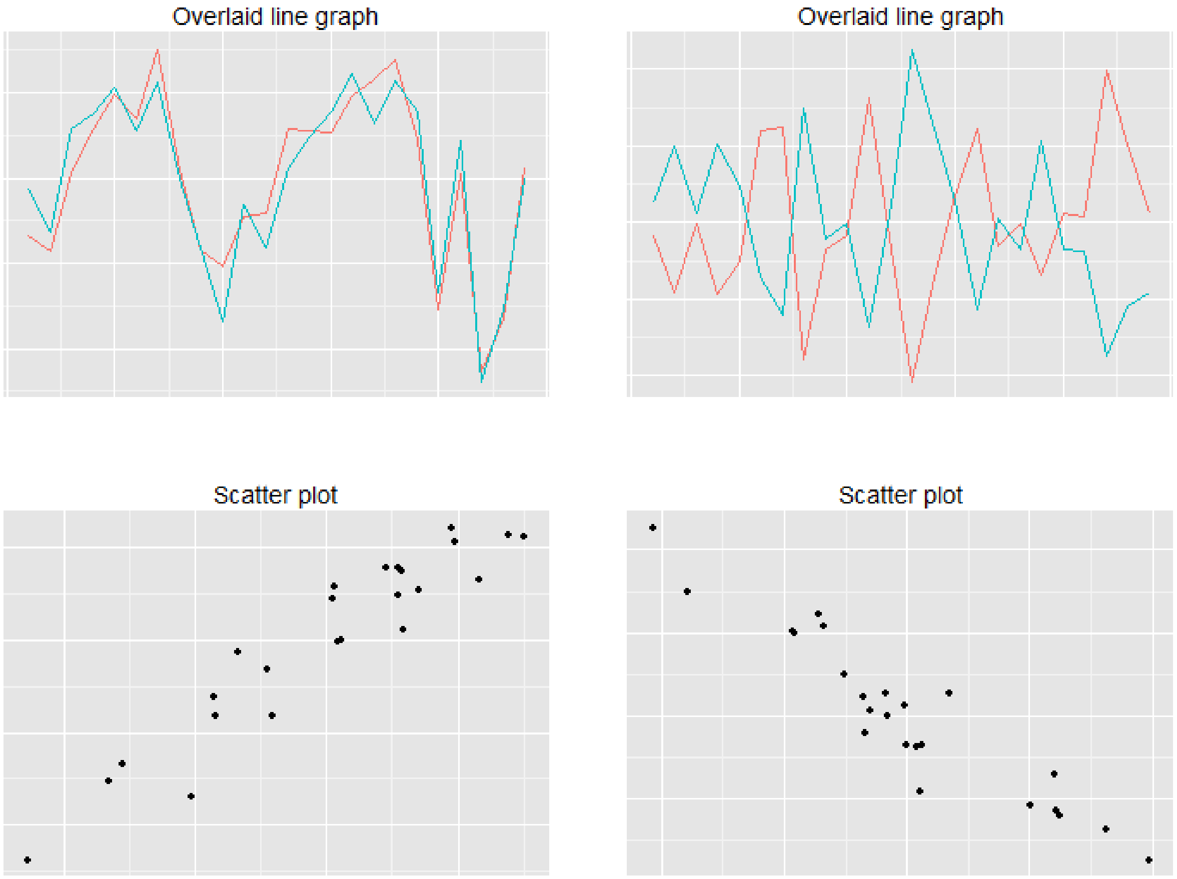


Figure 6: The same two data sets in both plot designs. t=24, correlation (left) = 0.946, correlation (right) = -0.943

4. Results

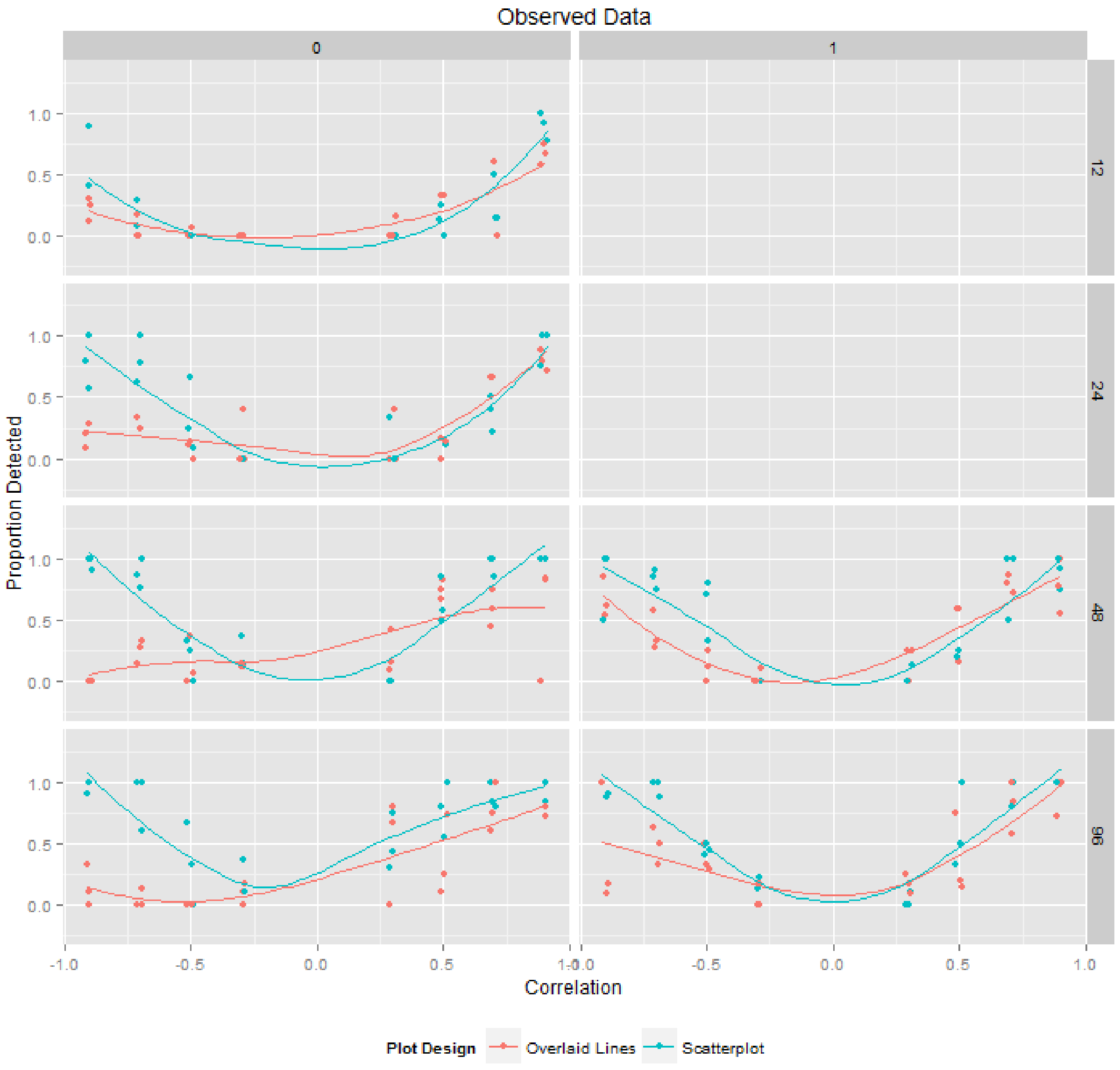


Figure 7: Response accuracy for unsmoothed data, broken by unsmoothed (left), smoothed (right) and sample size. The dots are data points, the lines represent a naive model average fit. Scatter plots generally showed superior performance, with the overlaid line graphs being particularly weak at revealing negative correlation

Figure 7 shows the raw results of the survey and figure 8 shows the difference in power with a 95% confidence interval calculated according to the formula in (2). Values above zero in figure 10 indicate that the scatter plot had more power to detect the true plot than the overlaid lines.

Performance between the two plot designs appears to be similar for positive correlation, with the scatter plot power gradually improving relative to the overlaid lines as sample size increases. The scatter plot appears to be the superior choice as it retains its power with negative correlation, where the overlaid lines performance suffered. This is not surprising, as positive and negative correlation appear as symmetric graphs in the scatter plot but the two characteristics appear very differently in the overlaid line graph. (See figure 6) The right side shows the performance benefits of smoothing overlaid line time series data. Much of the 'noise' in the time series is removed and studying the overall trend shows a greatly improved ability to successfully detect the true data plot when it is negatively correlated. The scatter plot still appears to be the superior plot, but much of the gap between the two is reduced for positive correlation.

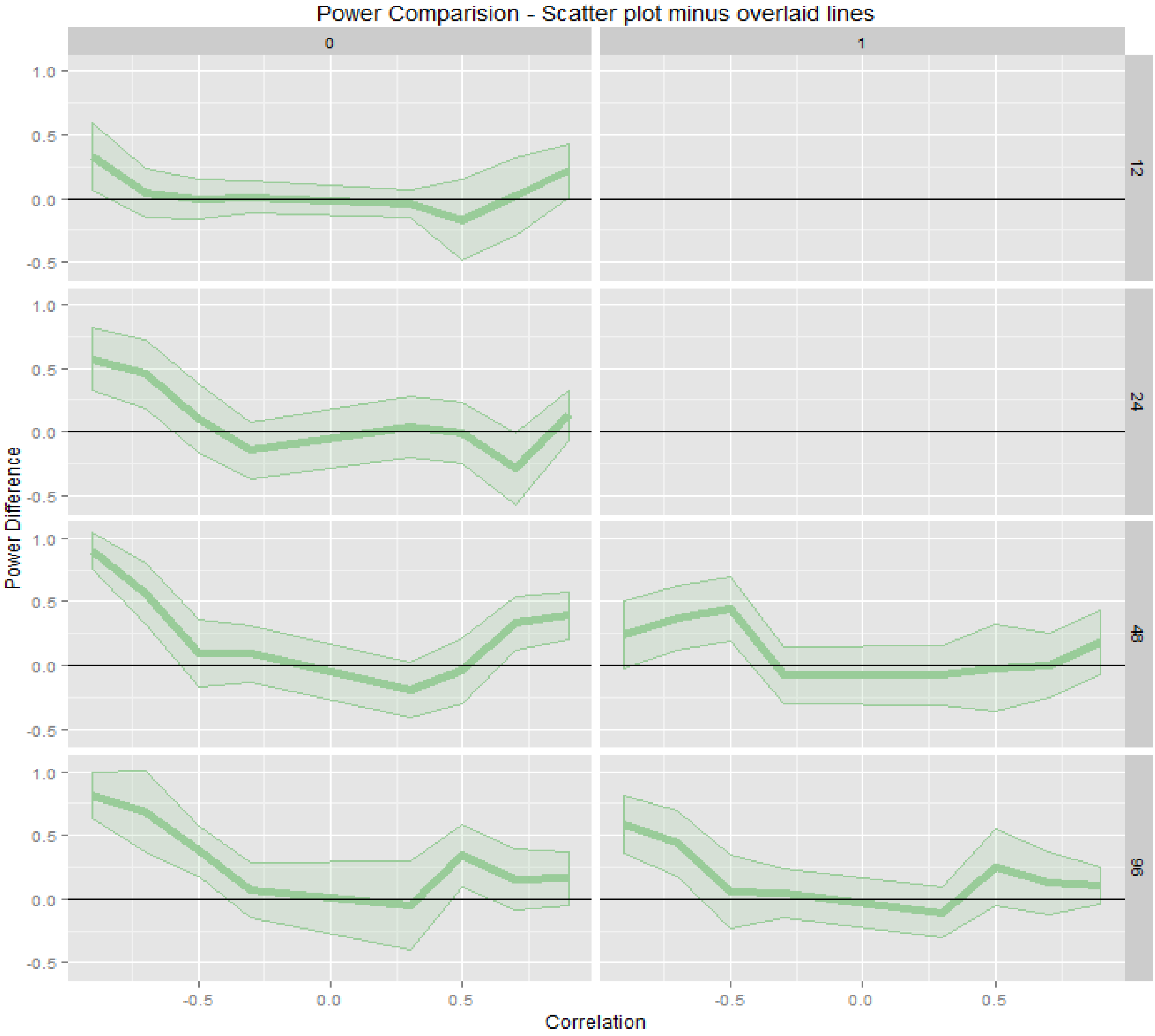


Figure 8: Difference in estimated power to reveal the true plot by smoothness and sample size with a 95% confidence interval. Values above zero indicate the scatter plot is more powerful than the overlaid line graph.

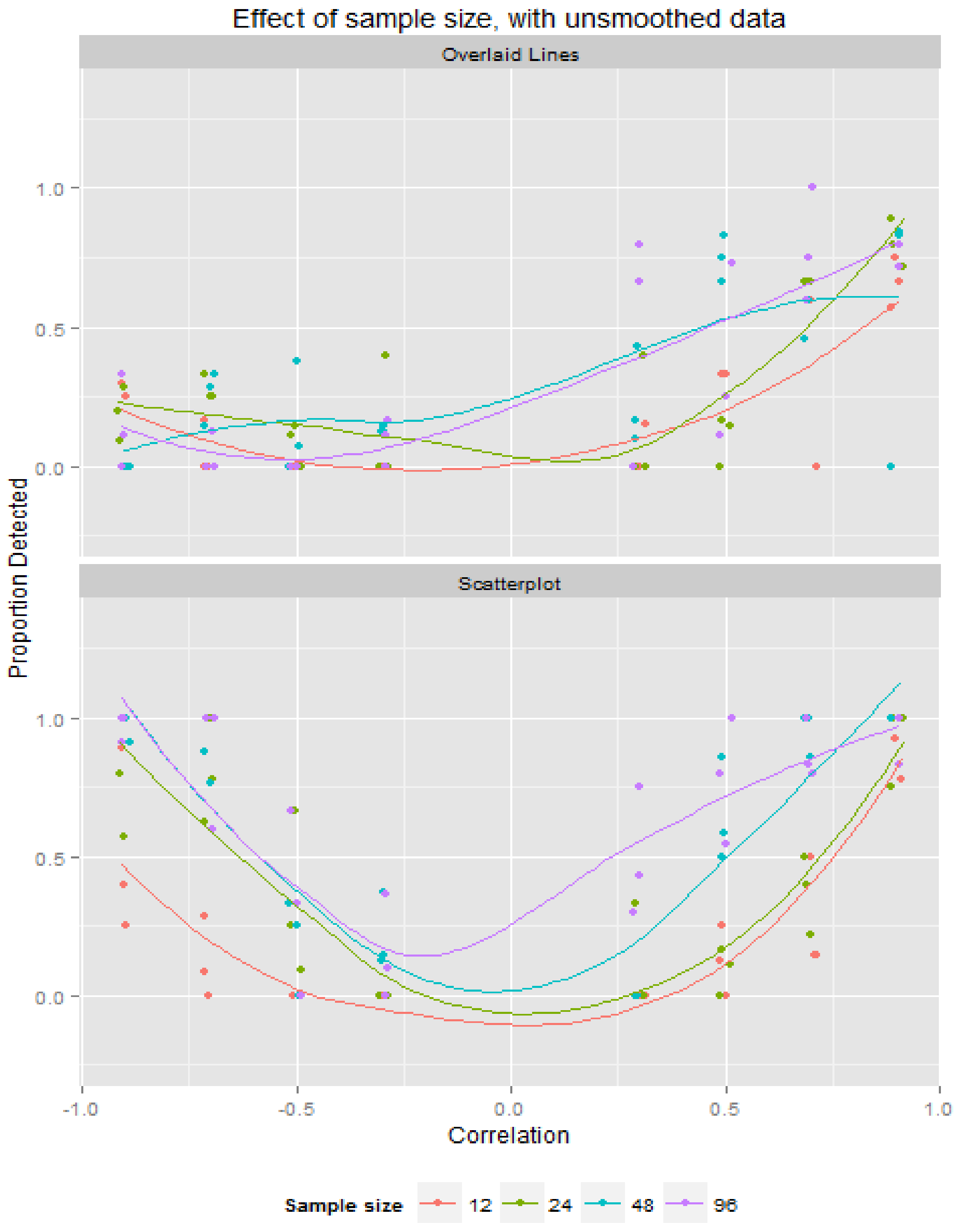


Figure 9: Effects of increasing sample size for unsmoothed data. Both plot designs appear to benefit from increased sample sizes.

Figure 9 demonstrates the effects of changing sample size for unsmoothed data. The scatter plot has increased power to detect correlation almost universally for any increase. Sample size influences the overlaid lines differently, they appears to benefit for positively correlated true plots, but it does not improve the weak performance in negative correlation.

Figures 10 and 11 give a breakdown of the interactions between the response variables: plot detection, time and confidence. Figure 8 is the histogram of the time taken to record a response, split by plot design and sign of correlation and true plot detection. The time data was highly positively skewed with a maximum of 1760 seconds so the variable was log-transformed. We found that the subject could identify the true plot out of the overlaid line graph 33.6% of the time and it took on average $e^{(3.442)}$ = 31.2 seconds to make a choice. Successful detections were slightly faster than unsuccessful decision, taking on average 28.7 seconds to pick the true plot and 32.6 seconds if a null was selected. The scatter plots performed much better, with the true plot picked 52.9% of the time, with an average of 22.7 seconds required to make a decision. If the true plot was detected it took only 18.6 seconds, and 29.3 seconds for a null plot selection.

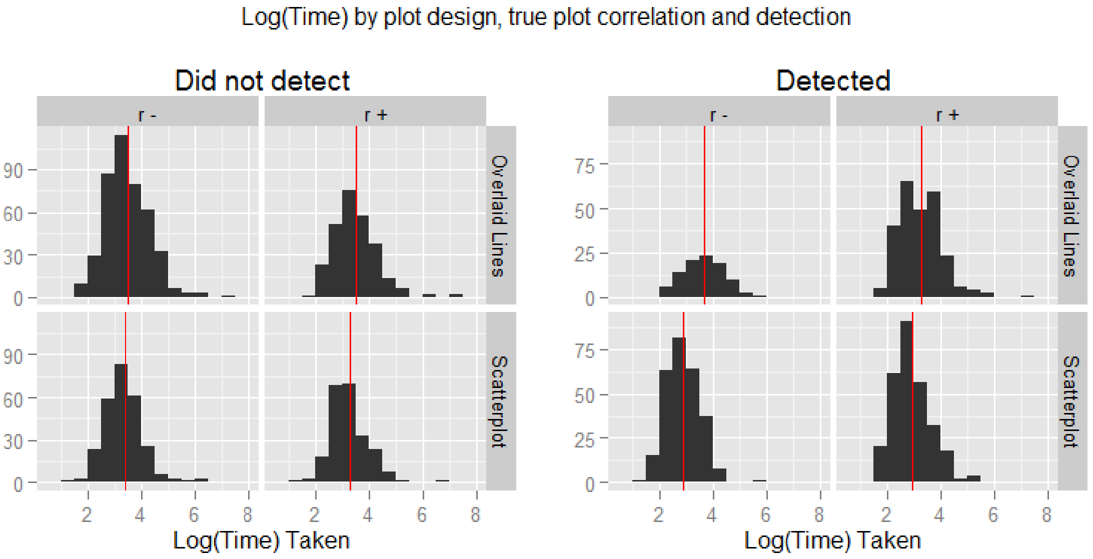


Figure 10: Histograms of log time taken to make a decision. r + and r - indicate the sign of true plot correlation. The vertical red line is the average time of the particular group. On average scatter plots are faster and the true plot is detected more often

Subjects were asked to give a number from 1 to 5 to rate how confident they were in there choice (1 - most confident, 5 - least confident, average = 3.34). There were no statistically significant differences between plot designs, however subjects were slightly less confident on true plot detections (by 0.25 ± 0.16, p=0.002) and every one percent extra time required to make a decision increased confidence (by 0.19 ± 0.07, p<0.001). One possible explanation is that subjects often focused on a null plot and picked it confidently. This would suggest that the null plots often produced visually interesting features subjects assumed were caused by the association between plots. .

Subjects spending longer on a plot immediately draws two hypotheses:

(a) The lineup was more difficult and it took longer to determine which plot was best, reducing confidence.

(b) The subject picked a plot relatively easily but spent extra time going over its details and making comparisons to alternative plots, increasing confidence.

The result suggests that hypothesis (b) is more likely to be correct. We also found a negative relationship between plot detection and time spent on a lineup, supporting the major relationship found: quick decisions are less confident but more likely to successfully detect the true plot.

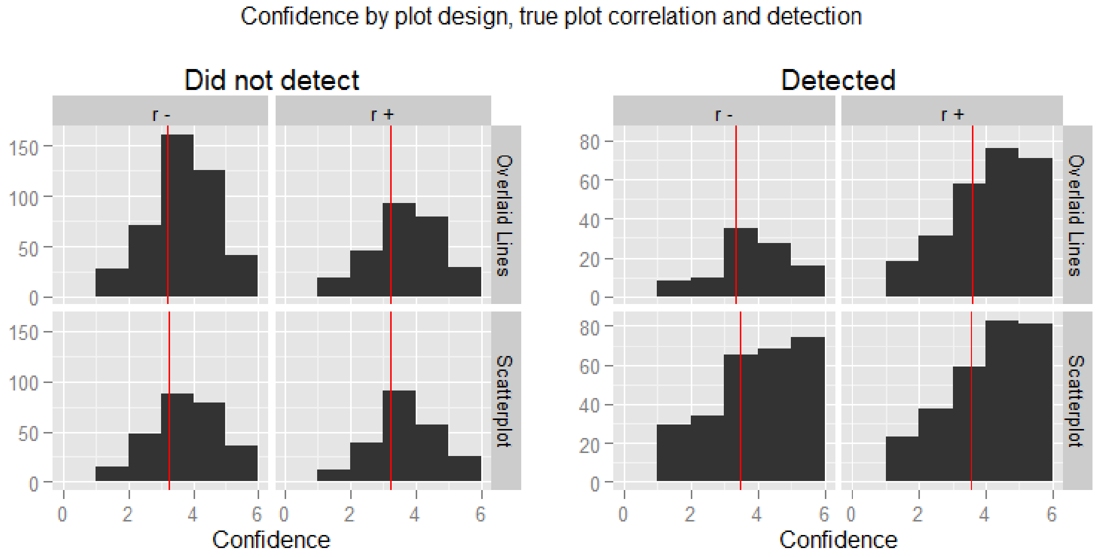


Figure 11: Histograms of reported confidence levels. 1 - Low, 5 - High. Correctness had a minor detrimental impact on confidence, whilst time taken increased it. There was no statistically significant difference on confidence level between plot designs.

Next we analysed the individual ability of each of our subjects. Figure 12 plots the accuracy of each subject on the y axis versus the average confidence on the x axis. While we expected that subjects with greater visual skills would report lower values of confidence, there does not appear to be any relationship between the two. The accuracy is normally distributed (Jarque-Bera p value = 0.418), with a mean of 42.5% and a standard deviation of 17.8%. The strongest subject successfully identified the true plot 91% of the time, whilst the weakest was not able to identify the true plot. As all subjects were shown a balanced mix of true plot correlations, the extreme difference in individual visual ability is striking. Visual ability will be controlled via a subject specific random effect in the model later in this section.

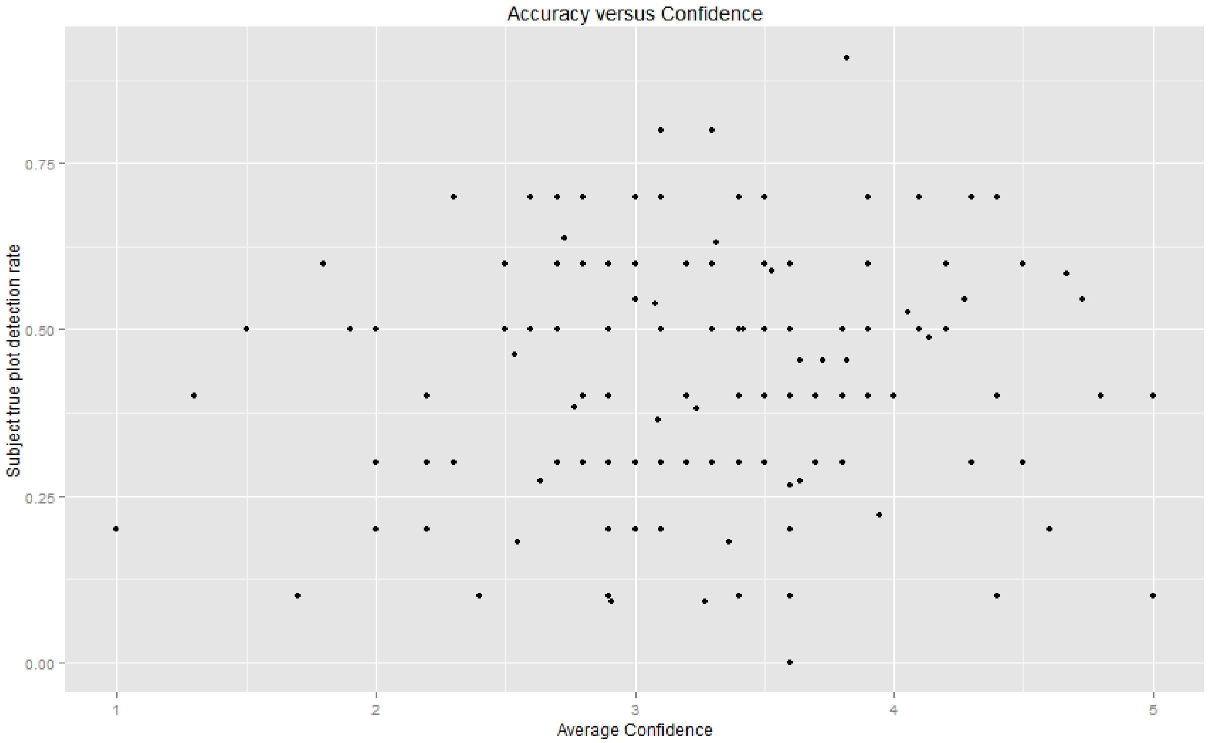


Figure 12: Individual subject accuracy versus confidence. There does not appear to be any relationship between the two.

There are two major reasons why a subject may fail to detect the true plot out of a lineup. Either the plot design was unable to display correlation for that particular combination of factors (and the analyst would commit a type 2 error by not rejecting the null hypothesis), or a null plot that had a greater correlation than the true plot was selected instead (and the analyst would commit a type 1 error by rejecting the null hypothesis). We next can analyse these kinds of errors with the pin plots found in figure 13. Figure 13 contains information on the frequency each plot, either true or null, was selected by an observer. The correlation of the selected plot is on the x-axis and the correlation of the simulated data in the plot is on the y-axis. The blue pins indicate the true plots and the red pins indicate the null plots. If the plot design has the power to display correlation, we would expect the highest correlation plot to be selected, leading to a Type 1 error where a plot from the null distribution spuriously had high correlation. We find that this is generally not the case. The selections appear to have a wide range of correlations, indicating that if a subject selected a null plot it was typically not due to excess spurious correlation. Instead the failure to select the highest correlated plot indicates that a Type 2 error was committed rather than Type 1 as the plot designs did not have the power to differentiate between medium and low correlation plots. However, for the scatter plot, whenever there was a plot with a very strong correlation, usually in the form of the true plot, it was overwhelmingly selected. For the overlaid lines this mostly occurred only for plots with strong positive correlation, which is consistent with its general lack of power to reveal negative relationships in data. Occasionally there was an anomaly where a single null plot was selected with a high frequency despite having a weak correlation. Occasionally the null distribution can

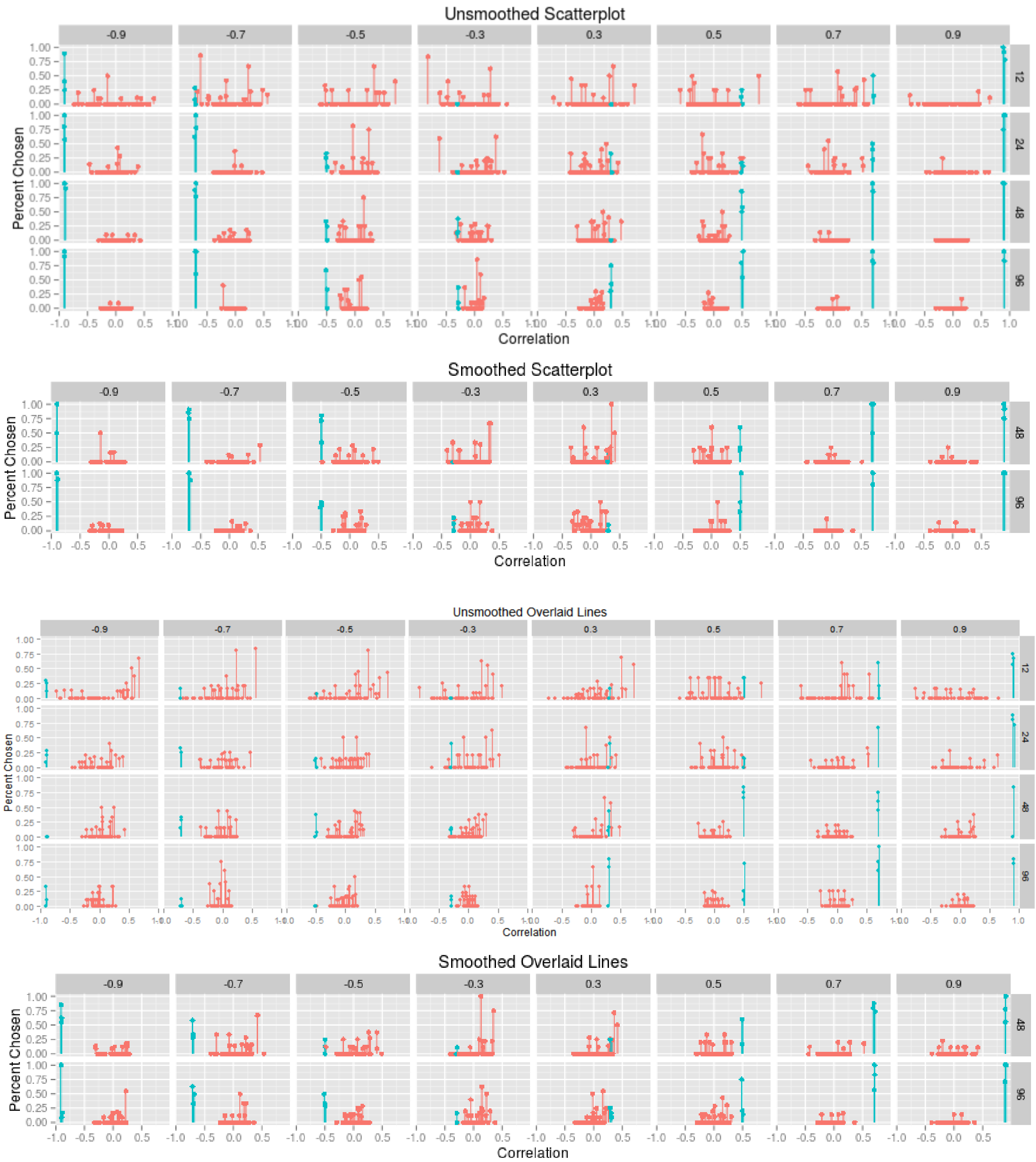


Figure 13: Pin plots of the subject selections of each lineup. On the x-axis is the correlation of the plot (there is potentially spurious correlation in the null plots), and on the y-axis is the proportion a plot was chosen. Red pins indicate null plots and blue indicates the true plot, occasionally the true plot does not appear as it has a frequency of zero and is visually blocked by the red pins. The pin plots are split with true plot correlation on the columns and sample size on the rows. Each plot contains information from all three lineups with that unique combination of factors: plot design, smoothness, sample size and true correlation.

generate visually interesting patterns that draw the analyst to confidently rejecting the null hypothesis. We asked subjects for reasons behind their selections found that choosing on the basis that the points in the plot were close together, or had a unique pattern more often led to an unsuccessful detection. Conversely, decisions based on matching the peaks and troughs of the overlaid lines or seeing a line formed in the dots of the scatter plot were more likely to be successful in identification of the true plot. However, these true detections are, on average, made less confidently than unsuccessful evaluations, which indicates that

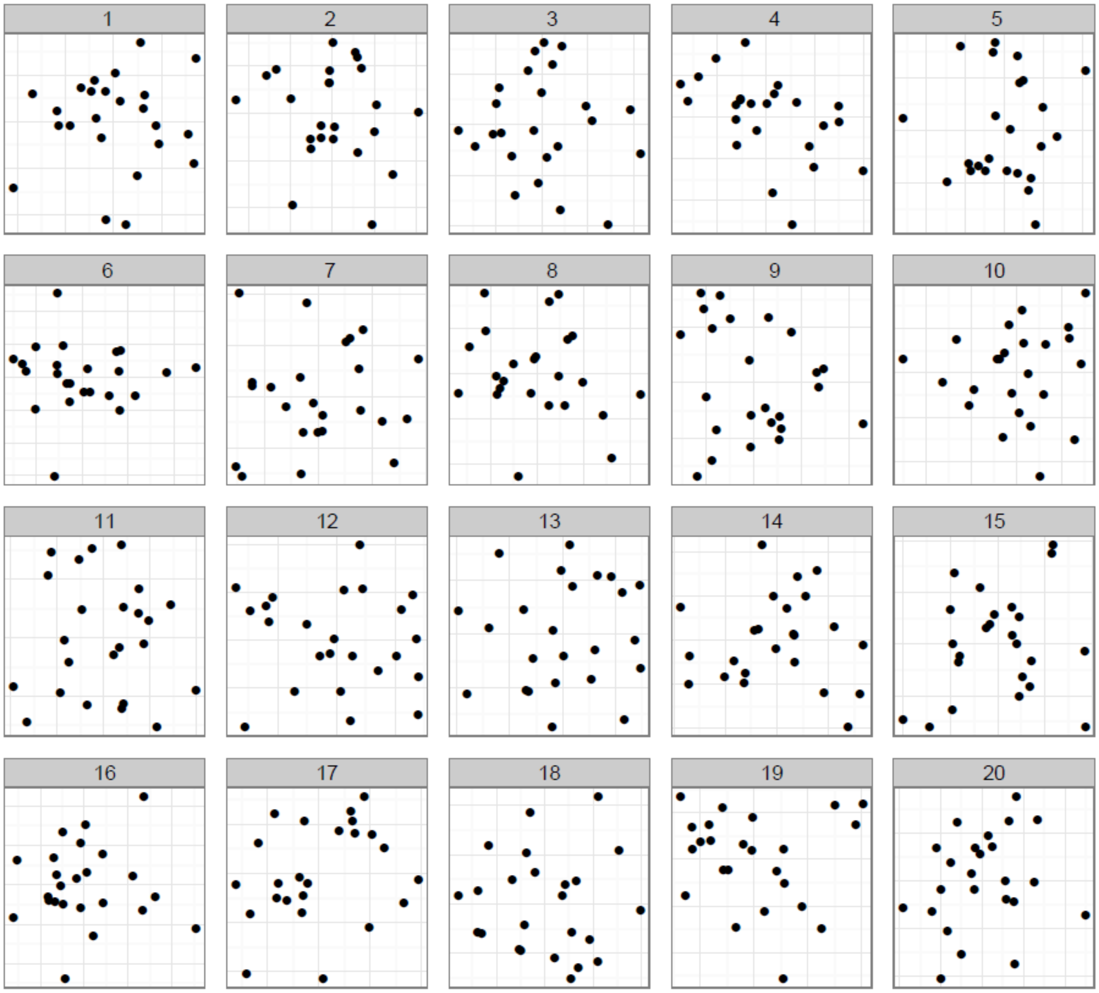


Figure 14: A lineup of unsmoothed scatter plots with a sample size of 24. The true plot is located in position four with a correlation of -0.491, and was selected by one out of eleven observers (p = 0.33, see formula (1)). The plot in position six was selected by nine of these eleven observers (p < 0.001) despite having a correlation of -0.037. This null plot shows a unique clustering effect that draws in the observer despite not having any linear relationship in the data, an example of the over-interpretation of randomness. Subjects most often cited the clustering pattern formed as the reason for their selection. The final selection was position eight, with a correlation of -0.226 (p = 0.33).

decisions made on the basis of closeness or patterns are more convincing of an association than the more successful reasons. Figures 14 and 15 demonstrate convincing null plots wrongly selected on this basis. Null plots within the plots powerful range of correlations tended to be selected, changing plot design can increase power (and thus reduce Type 2 error). It does not appear to decrease the Type 1 error of spuriously correlated plots under the null distribution being selected. Figure 15 below is another example of subjects over-interpreting randomness. Four out of five evaluations selected the null plot in position 13, which has a correlation of 0.225. It appears to be strongly correlated in the beginning of the series but breaks apart towards the end. The selection implies that subjects selecting this plot focused on the joint path of the two series early in the plot and ignored the contradictory evidence of no relationship as they diverged later in the plot. This selection is in-line with Wertheimer's gestalt law of common fate.

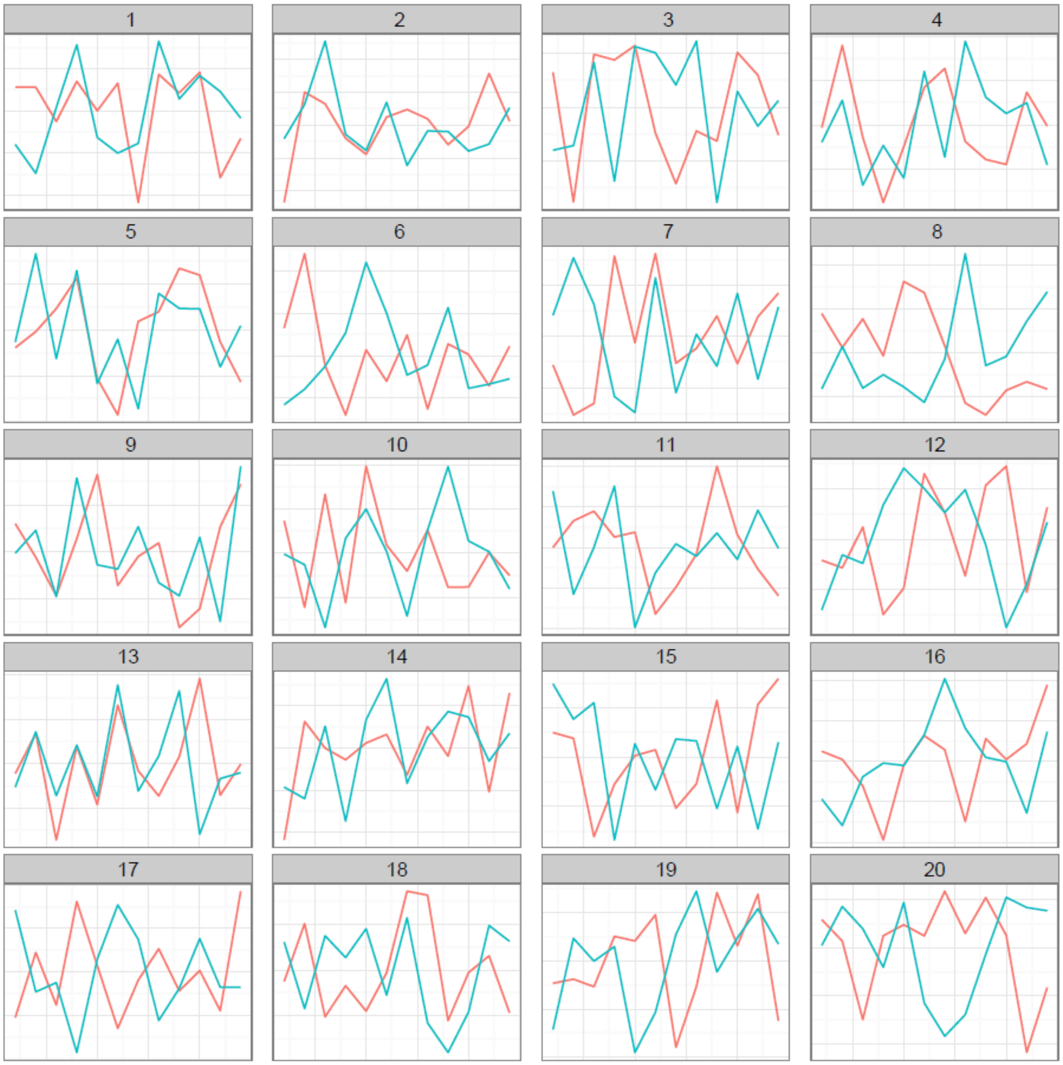


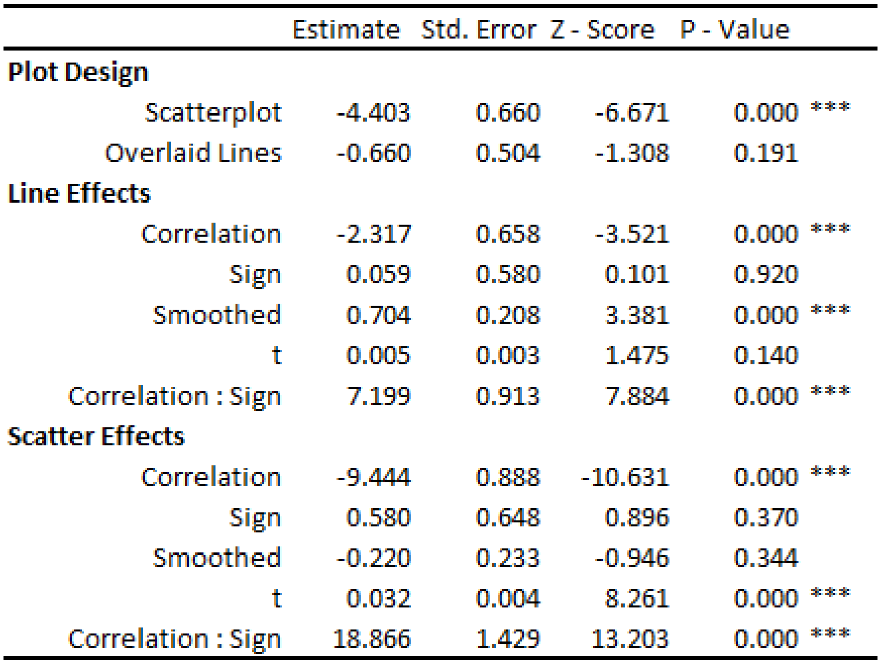
Figure 15: A lineup of unsmoothed overlaid lines with a sample size of 12 and true correlation of -0.707. The true plot in position 8 was not detected in five evaluations. However, plot 13 was selected four times, (p –value < 0.001) despite having a correlation of only 0.225. When asked for a reason of their selections, subjects answered that the lines were closest together.

A Mixed Effects Logit Model is used separately for scatter plot and overlaid line evaluations. The tested model is:

$$\mathrm{Power}_i = \psi\{\beta_{i,0} + \beta_{i,1} \cdot r + \beta_{i,2} \cdot \mathrm{Sign} + \beta_{i,3} \cdot \mathrm{Smoothed} + \beta_{i,4} \cdot t + \beta_{i,5} \cdot r \cdot \mathrm{Sign} + \gamma_{i,1} \cdot \mathrm{subject}_j\}$$

Where ψ is the logit link function, $\psi(x) = \frac{e^x}{1+e^x}$, *r* is the true plot correlation, Sign is the direction of the true plot correlation, *i* is the plot design and *j* is an observer specific random effect.

Table 1: Mixed Effects Logit Model for the predicted power of plot design. Sign = 1 if correlation > 0, 0 otherwise. Each plot design is estimated separately. Results must be transformed via the logit link function to generate a power prediction. Subjects with less than ten lineup evaluations were removed from the sample. Sample size for scatter plot model: 841, overlaid lines model: 843. The model is fitted using the lme4 package in R.



	Estimate	Std. Error	Z - Score	P - Value
Plot Design				
Scatterplot	-4.403	0.660	-6.671	0.000 ***
Overlaid Lines	-0.660	0.504	-1.308	0.191
Line Effects				
Correlation	-2.317	0.658	-3.521	0.000 ***
Sign	0.059	0.580	0.101	0.920
Smoothed	0.704	0.208	3.381	0.000 ***
t	0.005	0.003	1.475	0.140
Correlation : Sign	7.199	0.913	7.884	0.000 ***
Scatter Effects				
Correlation	-9.444	0.888	-10.631	0.000 ***
Sign	0.580	0.648	0.896	0.370
Smoothed	-0.220	0.233	-0.946	0.344
t	0.032	0.004	8.261	0.000 ***
Correlation : Sign	18.866	1.429	13.203	0.000 ***

*Signif codes: 0.00 ' *** ', 0.01 ' ** ', 0.05 ' * ', 0.1 '.'*

At first glance it appears the overlaid lines are substantially more powerful than the scatter plot, the interactions with correlation make this misleading. Taking correlation into account, the scatter plot shows typically superior performance, though the overlaid line graphs have a slight edge at extremely low values of correlation. However, we only tested correlations at ± 0.3, 0.5, 0.7 and 0.9, so the results below an absolute value of 0.3 are unreliable and may be artefacts of regression smoothing methods. The correlation:sign interaction shows that the scatter plot has close to symmetric performance for positive (shown by correlation:sign + correlation) and negative (Shown by -correlation) correlation while the overlaid line graph is substantially weaker for negative values of correlation. Surprisingly, smoothing out the excess 'noise' from a time series had a negative effect on scatter plots, despite a minimal visual impact of smoothed (See figure 5). It did have a significant positive impact on the power of overlaid line graphs.

Increasing the sample size, t, improved the power of the scatter plot but was not statistically significant for the overlaid line graph.

Figure 16 shows the fitted unrestricted model, with power predictions (y axis) plotted against correlation (x axis) for each of the four sample sizes tested. The thicker lines represent the fixed effect fit, or the average performance across all tested individuals whilst the thinner lines represent the random effects fit, the variation in power that can be attributed to an individual's visual skill. Previous studies [15] show that this random visual skill is not impacted by demographics such as age and gender; whilst education increases the power of a graphic display by only approximately one percent, not a practically significant amount. We did not include these characteristics of the individual in the modelling due to this lack of significance. Figure 16 especially highlights the weak performance of the overlaid lines and the benefits of smoothing.

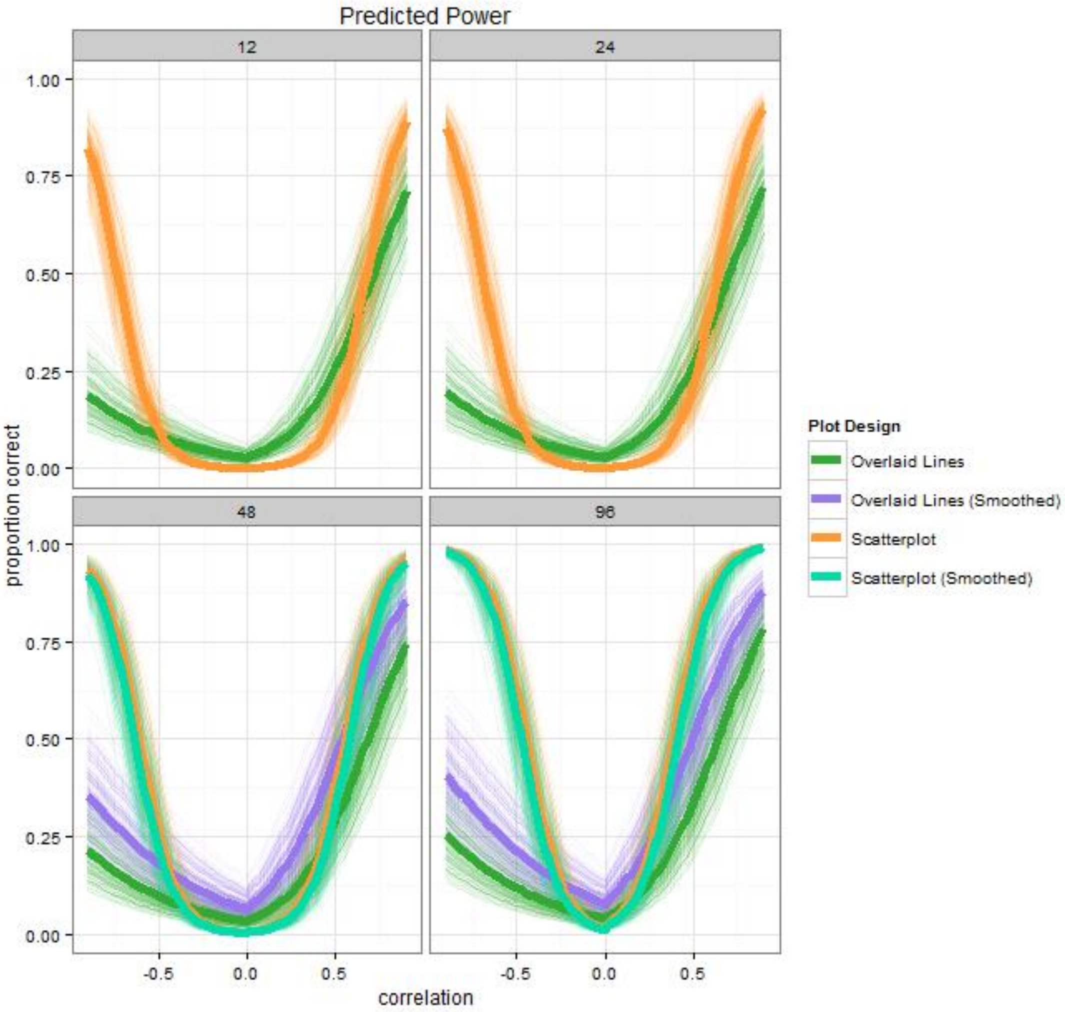


Figure 16: Predicted power fit from the mixed effects logit power model. The weakness of the overlaid line graph in negative correlation and the scatter plot's gain from sample size are highlighted.

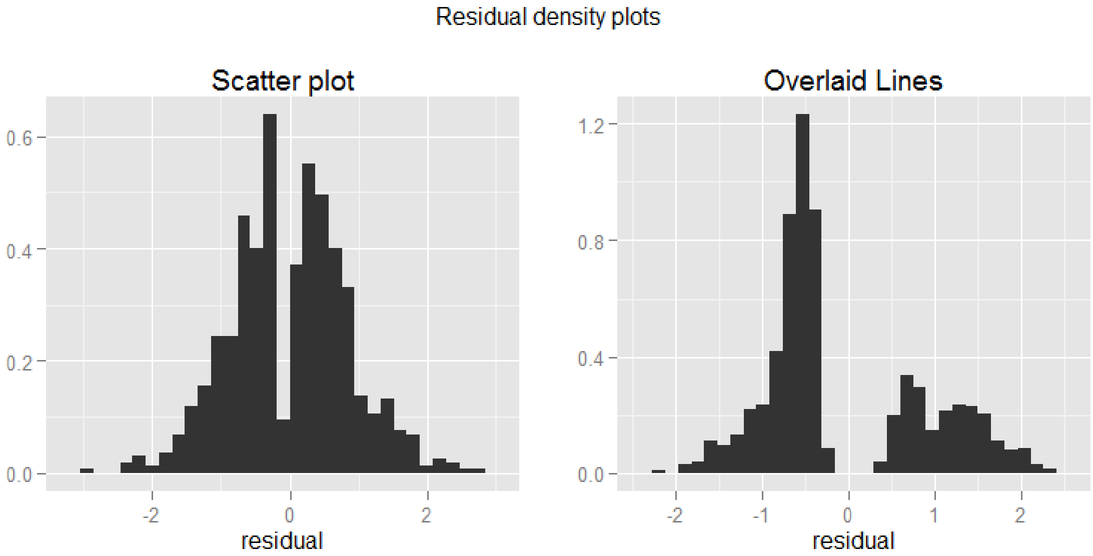


Figure 17: Histograms of residuals for both the scatter plot (left) and the overlaid lines (right).

Analysing the residuals allows us to test the model fit. As the dependent variable is a binary variable, the model will only have a residual close to zero if it predicts either a power of zero or one. As the rate of picking the true plot purely by chance is 1/m, or 0.05, the model should never predict a power of zero even if the plot has no ability to display correlation. This effect is especially notable in the overlaid line residual plot, the model fit is never close to zero or one so the residuals are split with no values around zero.

The scatter plot residuals are normally distributed with a Jarque-Bera normality test statistic of 0.09 (p-value = 0.95), however the overlaid lines strongly reject normality with a Jarque-Bera statistic of 75.6 (p-value = 0.00). The significance tests associated with the overlaid lines are only valid asymptotically. Also concerning, the mean value of the overlaid lines residuals is - 0.08, implicating that the model tends to over-predict power, further weakening the performance of this plot design relative to the scatter plot.

This unrestricted model used in to test the symmetry restriction for the scatter plot design is written formally as:

$$\text{Power}_{\text{UR},s} = \psi\{\beta_{s,0} + \beta_{s,1} \cdot |r| + \beta_{s,2} \cdot \text{Sign} + \beta_{s,3} \cdot \text{Smoothed} + \beta_{s,4} \cdot t + \beta_{s,5} \cdot |r| \cdot \text{Sign} + \gamma_{s,1} \cdot \text{subject}_j\}$$

Testing if the scatter plot displays positive and negative correlation equally is equivalent to testing the hypothesis:

$H_0: \beta_{s,2} = 0 \; \& \; \beta_{s.5} = 0$

H_1: Either equality does not hold.

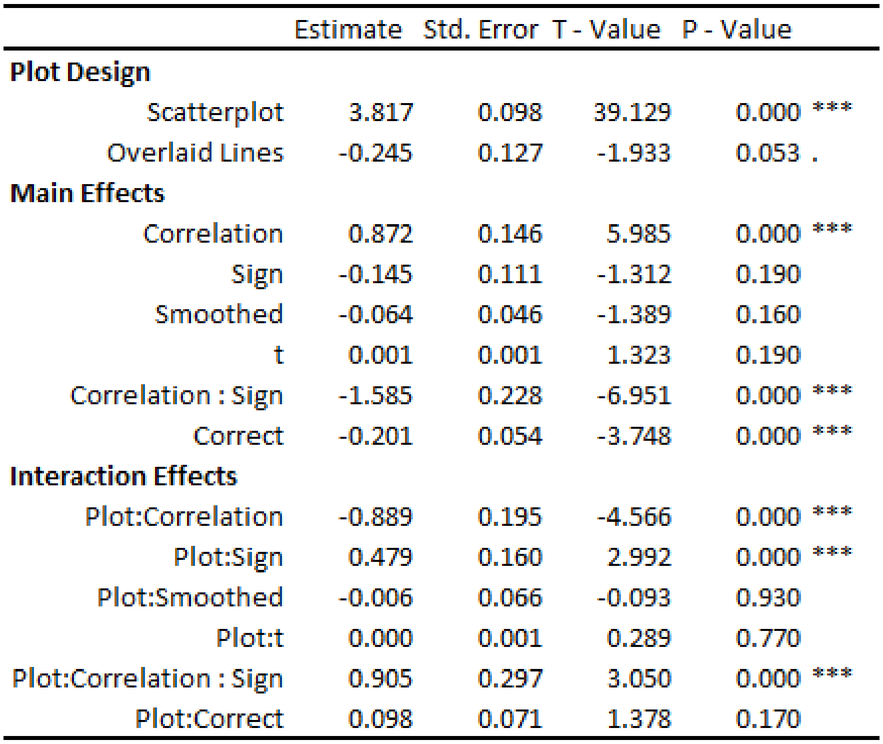
The restricted model is therefore:

$$\text{Power}_{\text{R}} = \psi\{\beta_{s,0} + \beta_{s,1} \cdot |\text{Correlation}| + \beta_{s,3} \cdot \text{Smoothed} + \beta_{s,4} \cdot t + \gamma_{s,1} \cdot \text{subject}_j\}$$

The unrestricted model has a log-likelihood value of -346.9, where the restricted model has a log-likelihood of -355.8. The test statistic is therefore $2(355.98 - 352.10) = 7.76$, with a null distributed as a χ_2^2 variable. The null hypothesis is rejected (p = 0.02). We find that the scatter plot has asymmetric performance over negative and positive values of correlation. This may be a result of unwittingly introducing bias into the methodology, the first scatter plot shown to subjects in the instruction set always had strong positive correlation.

A similar setup rejects the symmetric performance of the overlaid line graphs with a log likelihood ratio test statistic of 105.30 (p-value < 0.001). Such an extreme result is due to a systematic difference in the overlaid line graph and cannot be attributed simply to bias.

Table 2: Mixed Effects Linear Model with the dependent variable log(time). Sign = 1 if correlation > 0, 0 otherwise. Both plot designs estimated together with Overlaid Lines as the base design. Sample size: 1684. The model is fitted using the lme4 package in R.



	Estimate	Std. Error	T - Value	P - Value
Plot Design				
Scatterplot	3.817	0.098	39.129	0.000 ***
Overlaid Lines	-0.245	0.127	-1.933	0.053 .
Main Effects				
Correlation	0.872	0.146	5.985	0.000 ***
Sign	-0.145	0.111	-1.312	0.190
Smoothed	-0.064	0.046	-1.389	0.160
t	0.001	0.001	1.323	0.190
Correlation : Sign	-1.585	0.228	-6.951	0.000 ***
Correct	-0.201	0.054	-3.748	0.000 ***
Interaction Effects				
Plot:Correlation	-0.889	0.195	-4.566	0.000 ***
Plot:Sign	0.479	0.160	2.992	0.000 ***
Plot:Smoothed	-0.006	0.066	-0.093	0.930
Plot:t	0.000	0.001	0.289	0.770
Plot:Correlation : Sign	0.905	0.297	3.050	0.000 ***
Plot:Correct	0.098	0.071	1.378	0.170

*Signif codes: 0.00 ' *** ', 0.01 ' ** ', 0.05 ' * ', 0.1 '.'*

While it appears that the overlaid lines are on the verge of statistical significance (at the 5% level) for being the faster plot for a subject to visually process, this effect only holds for correlation values close to zero. The interactions between correlation and sign significantly reduce the time taken to pick the lineup for stronger correlations. Unsurprisingly, correct decisions are made faster. Surprisingly, neither smoothness nor sample size has an impact on the time for either plot design, despite having statistically significant impacts on power. An overlaid line graph having positive correlation increases the time required, eliminating the benefits from the almost significant plot design term. This result seems counter-intuitive as the weak performance of the overlaid lines in negative correlation implies that negative would require more time than positive, but the interactions between plot design, correlation and sign result in there being almost no decrease in time for strongly negatively correlated line plots.

5. Conclusion

This research showed that the scatterplot is better than the overlaid lines for displaying two time series when the purpose is to examine the association between the two series. This is especially true if correlation between the series is negative. This comparison was made possible by the use of the lineup protocol for comparing plot designs. The power of two plot designs for reading correlation between variables was modelled using a logistic regression incorporating subject's individual visual ability as a random effect. We found that the scatterplot, in most situations is both more powerful and quicker to process than the overlaid line graph. The overlaid line graph appeared to have stronger performance at extremely low correlations, but this research did not test absolute values of correlation below 0.3. Other non-linear associations may exist in data that are not measured by correlation, and it may be of interest to compare plot power with a broader range of true plot features.

Investigation of the individual selections of the lineup evaluations finds some results in-line with Wertheimer's Gestalt Law of common fate, as demonstrated in figure 15. The law suggested that people would find relationships between two lines that briefly move together, regardless of whether the relationship actually existed. Whilst this is shown, the overlaid line graph did not have a corresponding increase in confidence compared to the scatter plot; so research into similar confidence-boosting visual anomalies that can occur in scatter plots would be necessary to further investigate the law. Diaconis's Magical Thinking was supported by the finding on the relationships between time, confidence and correctness. A more difficult true plot may be picked rarely and require a longer to make a decision; resulting in decreased subject confidence. However, we found that time increases confidence but reduces correctness. Highly confident, but incorrect decisions, suggests subjects may have been distracted by strong, spurious patterns in the null plots.

If a data analyst is required to use overlaid line graphs in their work, substantial improvements to power and hence readability can be made by smoothing time series data if the sample size is long enough for these techniques to become feasible.

6. References

[1] Amazon.com Inc. 'Amazon Mechanical Turk', https://www.mturk.com/mturk/welcome, 2005, (accessed 29 April 2015).

[2] Buja, A. et al., 'Statistical inference for exploratory data analysis and model diagnostics', *Philosophical Transactions of the Royal Society*, vol. 367, no. 1906, 2009, pp 4361-4383.

[3] Cleveland, W. S., and R. McGill, 'Graphical Perception and Graphical Methods for Analyzing Scientific Data', *Science*, vol. 229, No. 4716, 1985, pp. 828-833.

[4] Daniel, C., 'Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments', Technometrics, vol. 1, no. 4, 1959, pp. 311-341.

[5] Dawson, K. et al., 'Two Graphical Techniques Useful in Detecting Correlation Structure in Repeated Measures Data', *The American Statistician*, vol. 51, no. 3, 1997, pp. 275-283.

[6] Diaconis, P., 'Magical Thinking in the Analysis of Scientific Data', *Annals of the New York Academy of Sciences*, vol. 364, no. 1, 1981, pp. 236-244

[7] Friendly, M., and D. J. Denis, 'Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualisation', http://datavis.ca/milestones/, 2001, (Accessed 1 May 2015).

[8] Harrison, L. et al, 'Ranking Visualizations of Correlation Using Weber's Law', *IEEE Conference on Information Visualization (Infovis)*, 2014.

[9] Heer, J., and M. Bostock, 'Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design', *CHI conference*, 2010.

[10] Hofmann, H. et al. 'Graphical Tests for Power Comparison of Competing Designs', *IEEE Transactions on Visualization & Computer Graphics*, vol. 18, no. 12, 2012, pp. 2441-2448.

[11] Hofmann, H. et al. 'Three scenarios of visual triangle tests', *Working paper*, Iowa State University, 2014.

[12] Javed, W. et al., 'Graphical Perception of Multiple Time Series', *IEEE Transactions on Visualization & Computer Graphics*, vol. 16, no. 6, 2010, pp. 927-934.

[13] Lee, S. H., and R. Blake, 'Visual Form Created Solely From Temporal Structure', *Science*, vol. 284, no. 5417, 1999, pp. 1165-1168.

[14] Li, J., et al. 'Judging correlation from scatterplots and parallel coordinate plots', *Information Visualization*, vol. 9, no. 1, 2010, pp. 13-30.

[15] Majumder, M. et al, 'Human Factors Influencing Visual Statistical Inference', *arXiv* 1408.1974, 2014

[16] Majumder, M. et al, 'Validation of Visual Statistical Inference, Applied to Linear Models', *Journal of the American Statistical Association*, vol. 108, no. 503, 2013, pp. 942-956.

[17] Nelson, C. R., and C. I. Plosser, 'Trends and Random Walks in Macroeconomic Time Series: Some evidence and implications', *Journal of Monetary Economics*, vol. 10, no 2, 1982, pp. 139-162.

[18] Perron, P., 'The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis', *Econometrica*, vol. 57, no. 6, 1989, pp. 1361-1401.

[19] Robbins, N., 'Plotting Differences: Don't Make Your Readers Do the Maths', http://www.forbes.com/sites/naomirobbins/2012/03/06/plotting-differences-dont-make-your-readers-do-the-math/, 2012, (accessed 1 May 2015).

[20] Sachs, J., '1786/1801: William Playfair, Statistical Graphics, and the Meaning of an Event', http://www.branchcollective.org/?ps_articles=jonathan-sachs-17861801-william-playfair-statistical-graphics-and-the-meaning-of-an-event, 2015, (accessed 29 April 2015).

[21] Sher, G., and P. Vitoria, 'An Information-Theoretic Test for Dependence with an Application to the Temporal Structure of Stock Returns', *Papers*, arXiv.org, 2013.

[22]Tuck, M., 'Gestalt Principles Applied in Design, http://sixrevisions.com/web_design/gestalt-principles-applied-in-design/, 2010, (accessed 29 April 2015).

[23] Tukey, J., 'The Technical Tools of Statistics', *The American Statistician*, vol. 19, no. 2, 1965, pp. 23-28.

[24] Turtle, H. J., 'Temporal Dependence in asset pricing models', *Economics Letters*, vol. 45, no. 1, 1994, pp. 361-366.

[25] VanderPlas, S., 'The perception of statistical graphics', PhD Thesis, Iowa State University, 2015.

[26] Wertheimer, M., 'Untersuchungen zur Lehre von der Gestalt II',*Psycologische Forschung*, vol. 4, 1923, pp 301-350, trans. W. Ellis, *A source book of Gestalt psychology*, London, Routledge & Kegan Paul, 1938, pp. 71-88.

[27] Wickham, H. et al., 'Graphical Inference for Infovis', *IEEE Transactions on Visualization & Computer Graphics*, vol. 16, no. 6, 2010, pp. 973-979.

Major software packages used:

R Core Team (2015). ‘R: A language and environment for statistical computing’, *R Foundation for Statistical Computing, Vienna Austria.*
http://www.R-project.org/.

Wickham, H. et al. ‘nullabor: Tools for Graphical Inference’ *R package version 0.3.1*.
http://CRAN.R-project.org/package=nullabor/.

Bates, D. et al. ‘lme4: Linear mixed-effects models using Eigen and S4’,
R package version 1.1-9.
http://CRAN.R-project.org/package=lme4/.

7. Appendices

A) Subjects

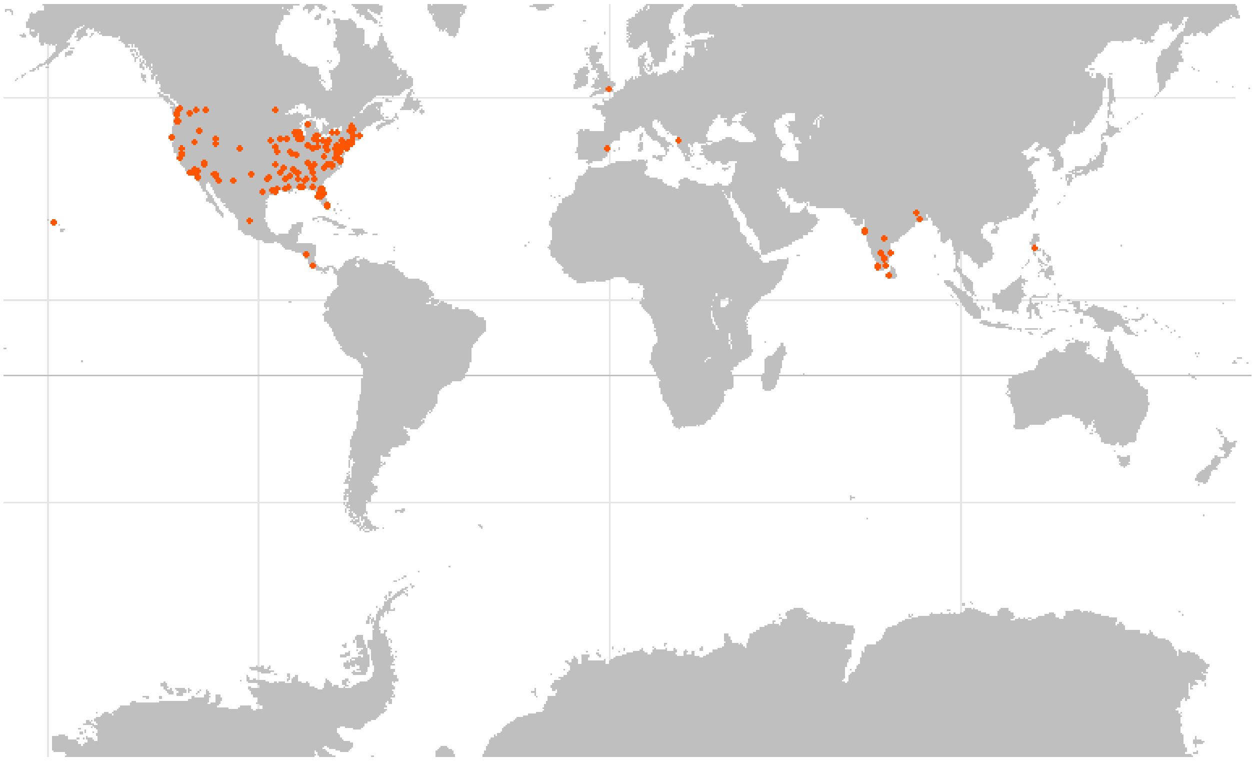


Figure 17: Geographical location of the observers. Most reside in the United States, but several are from Sri Lanka, India, Nicaragua, Spain, Costa Rica, Mexico, the Phillipines, the United Kingdom and Albania.

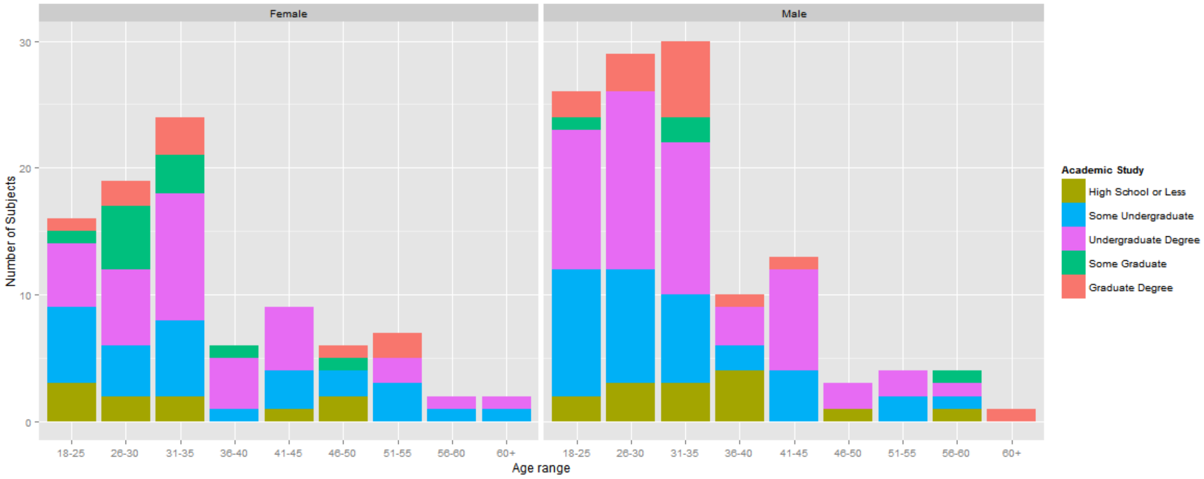
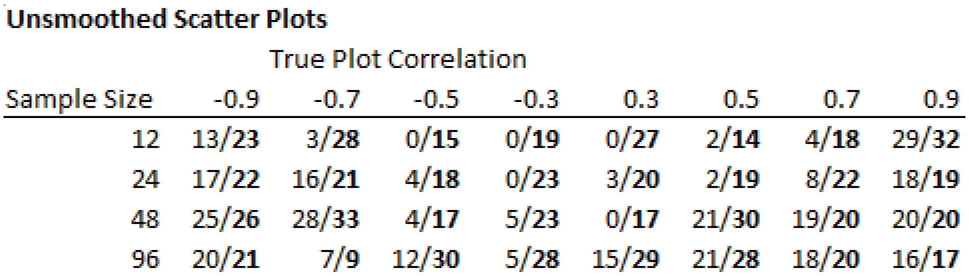


Figure 18: Users by gender, age range and level of academic study. Most of the users are in the younger age brackets but there is a balanced mixture of gender and education.

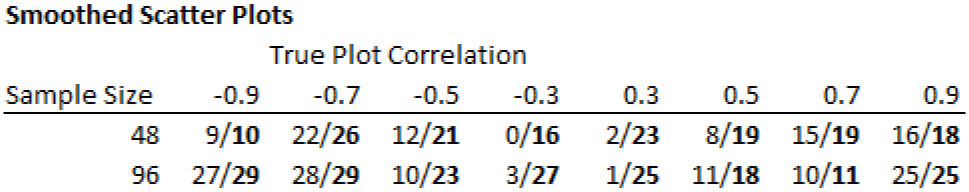
B) Tables of true plot detections

Table 3: The number of true plot detections/**number of lineup evaluations**



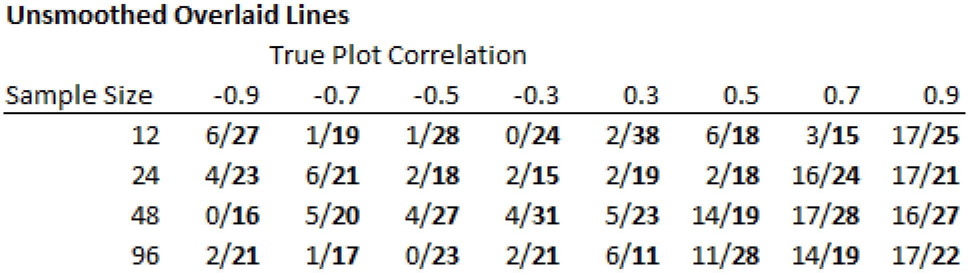
Unsmoothed Scatter Plots

	True Plot Correlation							
Sample Size	-0.9	-0.7	-0.5	-0.3	0.3	0.5	0.7	0.9
12	13/**23**	3/**28**	0/**15**	0/**19**	0/**27**	2/**14**	4/**18**	29/**32**
24	17/**22**	16/**21**	4/**18**	0/**23**	3/**20**	2/**19**	8/**22**	18/**19**
48	25/**26**	28/**33**	4/**17**	5/**23**	0/**17**	21/**30**	19/**20**	20/**20**
96	20/**21**	7/**9**	12/**30**	5/**28**	15/**29**	21/**28**	18/**20**	16/**17**



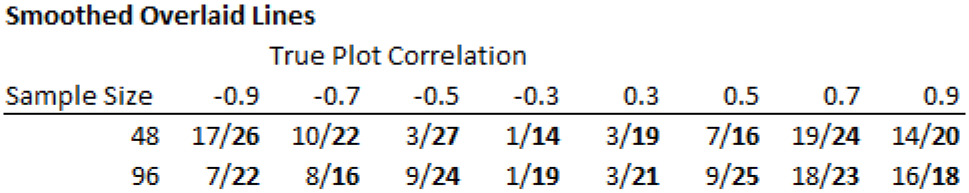
Smoothed Scatter Plots

	True Plot Correlation							
Sample Size	-0.9	-0.7	-0.5	-0.3	0.3	0.5	0.7	0.9
48	9/**10**	22/**26**	12/**21**	0/**16**	2/**23**	8/**19**	15/**19**	16/**18**
96	27/**29**	28/**29**	10/**23**	3/**27**	1/**25**	11/**18**	10/**11**	25/**25**



Unsmoothed Overlaid Lines

	True Plot Correlation							
Sample Size	-0.9	-0.7	-0.5	-0.3	0.3	0.5	0.7	0.9
12	6/**27**	1/**19**	1/**28**	0/**24**	2/**38**	6/**18**	3/**15**	17/**25**
24	4/**23**	6/**21**	2/**18**	2/**15**	2/**19**	2/**18**	16/**24**	17/**21**
48	0/**16**	5/**20**	4/**27**	4/**31**	5/**23**	14/**19**	17/**28**	16/**27**
96	2/**21**	1/**17**	0/**23**	2/**21**	6/**11**	11/**28**	14/**19**	17/**22**



Smoothed Overlaid Lines

	True Plot Correlation							
Sample Size	-0.9	-0.7	-0.5	-0.3	0.3	0.5	0.7	0.9
48	17/**26**	10/**22**	3/**27**	1/**14**	3/**19**	7/**16**	19/**24**	14/**20**
96	7/**22**	8/**16**	9/**24**	1/**19**	3/**21**	9/**25**	18/**23**	16/**18**